< 빅데이터분석 2021 기말 실기 고사 >

☞ **제출 파일 형식 예) 김한신_201858220_빅데이터기말실기.html**

☞ **이름, 학번 그리고 문제 번호, 질문에 대한 답을 적절한 위치에 마크다운으로 표시할 것**

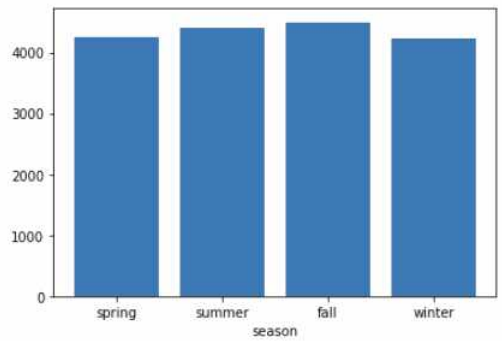
[A] 오늘의 기상 관측치들(c1~c15)로부터 다음날 비가 오는지(rain)를 예측하는 분석을 하고자 한다. 다음 순서대로 필요한 코드를 작성하고 질문에 답하시오. (5점*7문제=35점)

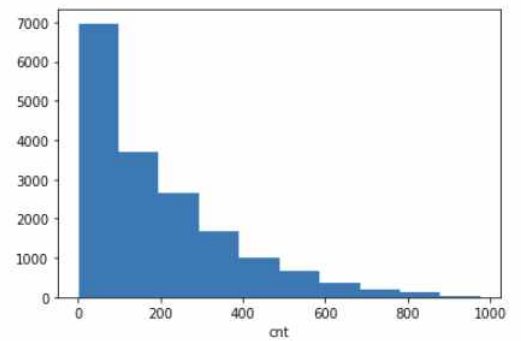
1) 'FinalA.txt' 파일을 읽어 DataFrame 객체를 생성하여 처음 5행을 출력하고, 샘플의 개수를 구하시오.
2) 특성 행렬과 레이블 벡터(단, 'Yes'는 1, 'No'는 0으로 코딩)를 생성하시오.
3) 각 특성의 평균과 레이블의 분포(값별 빈도수)를 구한 후, 데이터 표준화와 데이터 분할 시 층화추출(stratify)이 필요한 이유를 설명하시오.
4) 데이터를 8:2로 훈련용과 테스트용으로 분할하고(단, stratify=y, random_state=0) 표준화하시오.
5) 로지스틱회귀(디폴트모형)를 실시하여 테스트 데이터에 대한 정확도와 정오분류표를 출력하고 'Yes' 클래스의 정밀도와 재현율을 구하시오.
6) max_depth 값을 1부터 10까지 변화시키면서 교차검증(cv=5)을 통하여 최적 max_depth를 구하여, 이 값을 사용하는 결정트리에 사용되는 특성들의 이름을 구하시오.
7) 은닉층이 한 개(유닛 개수=16)인 MLP 신경망을 구축하고 에포크수를 10, 배치사이즈를 2로 하여 훈련시킨 후, 테스트셋의 정확도를 구하시오.

[B] 날짜와 기후 특성을 사용하여 자전거 공유시스템 사용 건수(cnt)를 예측하는 분석을 하고자 한다. 다음 순서대로 필요한 코드를 작성하고 질문에 답하시오. (5점*4문제=20점)

칼럼	변수명	설명	코딩 방식
0	season	계절	1:봄, 2:여름, 3:가을, 4:겨울
1	weekday	요일	0:일, 1:월, ..., 6:토
2	weathersit	날씨 구분	1:맑음/약간흐림 2:안개/흐림 3:약한눈/비 4:강한눈/비
3	temp	온도	0~1로 표준화 됨
4	hum	습도	0~1로 표준화 됨
5	windspeed	풍속	0~1로 표준화 됨
6	cnt	사용자 수	레이블

1) 'FinalB.txt' 파일을 읽어 DataFrame을 생성한 후, 숫자로 인코딩된 범주형 칼럼의 타입을 예와 같이 변경하여 info() 메서드 결과를 출력하시오. [예] df['season'] = df['season'].astype(str)
2) season 칼럼과 cnt 칼럼에 대하여 아래와 같은 그래프를 작성하시오.





3) temp, hum, windspeed를 특성으로 하여 cnt를 설명하는 선형회귀를 실시하고 결정계수를 구하시오. (답:0.25)
4) 모든 특성들을 사용하여(범주형 특성들은 원핫인코딩) cnt를 설명하는 선형회귀를 실시하고 결정계수를 구하시오. (답:0.28)